Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization

Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization *

Ying Zhang[1], Ömer Deniz Akyildiz[2,3], Theo Damoulas[2,3], and Sotirios Sabanis[1,3]

[1]School of Mathematics, The University of Edinburgh, UK.
[2]The University of Warwick, UK.
[3]The Alan Turing Institute, UK.

Abstract

Within the context of empirical risk minimization, see Raginsky, Rakhlin, and Telgarsky (2017), we are concerned with a non-asymptotic analysis of sampling algorithms used in optimization. In particular, we obtain non-asymptotic error bounds for a popular class of algorithms called Stochastic Gradient Langevin Dynamics (SGLD). These results are derived in appropriate Wasserstein distances in the absence of the log-concavity of the target distribution. More precisely, the local Lipschitzness of the stochastic gradient $H(\theta, x)$ is assumed, and furthermore, the dissipativity and convexity at infinity condition are relaxed by removing the uniform dependence in x.

1 Introduction

Consider a non-convex stochastic optimization problem

$$\text{minimize} \quad U(\theta) := \mathbb{E}[f(\theta, X)],$$

where $\theta \in \mathbb{R}^d$ and X is a random element with some unknown probability law. We aim to generate $\hat{\theta}$ such that the expected excess risk

$$\mathbb{E}[U(\hat{\theta})] - \inf_{\theta \in \mathbb{R}^d} U(\theta)$$

is minimized. The optimization problem of minimizing U can be decomposed into three subproblems as explained in [15], and one of which is a problem of sampling from the target distribution $\pi_\beta(\theta) \propto \exp(-\beta U(\theta))$. Intuitively, the two problems are linked since π_β concentrates around the minimizers of U when $\beta > 0$ takes sufficiently large values (see [13]). Moreover, it is well-known that, under mild conditions, the Langevin SDE associated with π_β is given by

$$\mathrm{d}Z_t = -h(Z_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}B_t \tag{1}$$

with a (possibly random) initial condition θ_0, where $h := \nabla U$ and $(B_t)_{t \geq 0}$ is a d-dimensional Brownian motion. A standard approach to sample from the target distribution π_β is to approximate the Langevin SDE (1) by using an Euler discretization scheme which serves as a sampling algorithm, known as the unadjusted Langevin algorithm (ULA) or Langevin Monte Carlo (LMC) algorithm. Theoretical guarantees for the convergence of ULA in Wasserstein distance and in total variation have been obtained under the assumption that U is strongly convex with globally Lipschitz gradient, see [7],

*This work was supported by The Alan Turing Institute for Data Science and AI under EPSRC grant EP/N510129/1. Ö. D. A. and T. D. are funded by the Lloyds Register Foundation programme on Data Centric Engineering through the London Air Quality project. Y. Z. was supported by The Maxwell Institute Graduate School in Analysis and its Applications, a Centre for Doctoral Training funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016508/01), the Scottish Funding Council, Heriot-Watt University and the University of Edinburgh.

[9] and [10]. Extensions which includes locally Lipschitz gradient and higher order algorithms can be found in [3], [8] and [17].

In practice, however, the gradient h is usually unknown and one only has an unbiased estimate of h. A natural extension of ULA, which was introduced in [16] in the context of Bayesian inference and which has found great applicability in this type of stochastic optimization problems, is the Stochastic Gradient Langevin Dynamics (SGLD) algorithm. More precisely, fix an $\mathbb{R}^d$-valued random variable θ_0 representing its initial value and let $(X_n)_{n\in\mathbb{N}}$ be an i.i.d. sequence, the SGLD algorithm corresponding to SDE (1) is given by,

$$\theta_0^\lambda := \theta_0, \quad \theta_{n+1}^\lambda := \theta_n^\lambda - \lambda H(\theta_n^\lambda, X_{n+1}) + \sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}, \; n \in \mathbb{N}, \tag{2}$$

where $\lambda > 0$ is often called the stepsize or gain of the algorithm, $\beta > 0$ is the so-called inverse temperature parameter, $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is a measurable function satisfying $h(\theta) = \mathbb{E}[H(\theta, X_0)]$ and $(\xi_n)_{n\in\mathbb{N}}$ is an independent sequence of standard d-dimensional Gaussian random variables. The properties of the i.i.d. process $(X_n)_{n\in\mathbb{N}}$ are given below. For a strongly convex stochastic gradient H, [1], [2] and [8] obtain non-asymptotic bounds in Wasserstein-2 distance between the SGLD algorithm and the target distribution π_β. While [8] assumes H is a linear combination of h and $(X_n)_{n\in\mathbb{N}}$, which allows bounded conditional bias, a general form of H with non-Markovian $(X_n)_{n\in\mathbb{N}}$ is considered in [1]. For the case where the convexity does not hold, it is a challenging task to obtain convergence results in Wasserstein distance. One line of research is to replace the convexity condition with a dissipativity condition. The first such non-asymptotic estimate is provided by [15] in Wasserstein-2 distance although its rate of convergence is $\lambda^{5/4}n$ which depends on the number of iteration. Recently, improved results were presented in [5], where the convergence rate $1/2$ in a bounded Wasserstein distance is obtained. The analysis in [5] relies on the construction of certain auxiliary continuous processes and the contraction results in [12]. Another line of research is to assume convexity at infinity. [6] and [14] obtain convergence results in Wasserstein-1 distance by using the contraction property developed in [11].

In this paper, we consider the case where H is locally Lipschitz continuous in both state variables as stated in Assumption 2 below. Our approach is motivated by popular applications in statistical and machine learning. Crucially, we relax substantially the assumptions of dissipativity and convexity at infinity on the stochastic gradient $H(\theta, x)$ by allowing non-uniform dependence in x. We establish in Theorem 2.3 and 2.5 non-asymptotic convergence results for the SGLD algorithm (2) in bounded Wasserstein distance and in Wasserstein-1 distance respectively.

We conclude this section by introducing some notation. Let $(\Omega, \mathcal{F}, P)$ be a probability space. We denote by $\mathbb{E}[X]$ the expectation of a random variable X. For $1 \le p < \infty$, L^p is used to denote the usual space of p-integrable real-valued random variables. Fix an integer $d \ge 1$. For an $\mathbb{R}^d$-valued random variable X, its law on $\mathcal{B}(\mathbb{R}^d)$ (the Borel sigma-algebra of $\mathbb{R}^d$) is denoted by $\mathcal{L}(X)$. Scalar product is denoted by $\langle\cdot,\cdot\rangle$, with $|\cdot|$ standing for the corresponding norm (where the dimension of the space may vary depending on the context). For $\mu \in \mathcal{P}(\mathbb{R}^d)$ and for a non-negative measurable $f : \mathbb{R}^d \to \mathbb{R}$, the notation $\mu(f) := \int_{\mathbb{R}^d} f(\theta)\mu(\mathrm{d}\theta)$ is used. For any integer $q \ge 1$, let $\mathcal{P}(\mathbb{R}^q)$ denote the set of probability measures on $\mathcal{B}(\mathbb{R}^q)$. For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, let $\mathcal{C}(\mu, \nu)$ denote the set of probability measures ζ on $\mathcal{B}(\mathbb{R}^{2d})$ such that its respective marginals are μ, ν. For two probability measures μ and ν, the Wasserstein distance of order $p \ge 1$ is defined as

$$\tilde{W}_p(\mu, \nu) := \inf_{\zeta\in\mathcal{C}(\mu,\nu)} \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\theta - \theta'|^p \zeta(\mathrm{d}\theta \mathrm{d}\theta') \right)^{1/p}, \; \mu, \nu \in \mathcal{P}(\mathbb{R}^d). \tag{3}$$

Then, define

$$W_1(\mu, \nu) := \inf_{\zeta\in\mathcal{C}(\mu,\nu)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} [|x - y| \wedge 1]\zeta(dx, dy), \tag{4}$$

which is the Wasserstein-1 distance associated to the bounded metric $|x - y| \wedge 1$, $x, y \in \mathbb{R}^d$.

2 Main results and comparisons

Denote by $\mathcal{G}_n := \sigma(X_k,\ k \le n,\ k \in \mathbb{N})$, for any $n \in \mathbb{N}$. $(X_n)_{n\in\mathbb{N}}$ is an $\mathbb{R}^m$-valued, $(\mathcal{G}_n)_{n\in\mathbb{N}}$-adapted process. We assume throughout the paper that θ_0, $\mathcal{G}_\infty$ and $(\xi_n)_{n\in\mathbb{N}}$ are independent. Then, we present the following assumptions:

Assumption 1. *The process $(X_n)_{n\in\mathbb{N}}$ is i.i.d. with $|X_0| \in L^{4(\rho+1)}$ and $|\theta_0| \in L^4$. It satisfies*

$$\mathbb{E}[H(\theta, X_0)] = h(\theta).$$

Assumption 2. *There exist positive constants L_1, L_2 and ρ such that, for all $x, x' \in \mathbb{R}^m$ and $\theta, \theta' \in \mathbb{R}^d$,*

$$\begin{aligned} |H(\theta, x) - H(\theta', x)| &\le L_1(1 + |x|)^\rho |\theta - \theta'|, \\ |H(\theta, x) - H(\theta, x')| &\le L_2(1 + |x| + |x'|)^\rho (1 + |\theta|)|x - x'|. \end{aligned}$$

We set

$$H_\star := |H(0, 0)|. \tag{5}$$

Remark 2.1. *Assumption 2 implies, for all $\theta, \theta' \in \mathbb{R}^d$,*

$$|h(\theta) - h(\theta')| \le L_1 \mathbb{E}[(1 + |X_0|)^\rho]|\theta - \theta'|\,. \tag{6}$$

Also, Assumption 2 implies

$$|H(\theta, x)| \le L_1(1 + |x|)^\rho |\theta| + L_2(1 + |x|)^{\rho+1} + H_\star.$$

Moreover, one notices that under Assumptions 1 and 2, $h(\theta) = \mathbb{E}[H(\theta, X_0)]$ for all $\theta \in \mathbb{R}^d$, is well-defined.

Assumption 3. *There exist $A : \mathbb{R}^m \to \mathbb{R}^{d\times d}$, $b : \mathbb{R}^m \to \mathbb{R}$ such that for any $x, y \in \mathbb{R}^d$,*

$$\langle y, A(x)y\rangle \ge 0$$

and for all $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$,

$$\langle H(\theta, x), \theta\rangle \ge \langle \theta, A(x)\theta\rangle - b(x).$$

The smallest eigenvalue of $\mathbb{E}[A(X_0)]$ is a positive real number $a > 0$ and $E[b(X_0)] = b > 0$.

Remark 2.2. *By Assumption 3, one obtains, for $\theta \in \mathbb{R}^d$*

$$\langle h(\theta), \theta\rangle \ge a|\theta|^2 - b.$$

Denote by

$$C_\rho = \mathbb{E}\left[(1 + |X_0|)^{4(\rho+1)}\right]. \tag{7}$$

Define

$$\lambda_{\max} = \min\left\{\frac{\min\{a, a^{1/3}\}}{36(1 + L_1)^2\left(\mathbb{E}\left[(1 + |X_0|)^{4\rho}\right]\right)^{1/2}}, \frac{1}{a}\right\}, \tag{8}$$

where L_1 and a are defined in Assumptions 2 and 3, respectively.

Theorem 2.3. *Let Assumptions 1, 2 and 3 hold. Then, there exist constants $C_0, C_1, C_2 > 0$ such that, for every $0 < \lambda \le \lambda_{\max}$,*

$$W_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \le C_1 e^{-C_0\lambda n}(\mathbb{E}[|\theta_0|^4] + 1) + C_2\sqrt{\lambda},\ n \in \mathbb{N}, \tag{9}$$

where C_0, C_1 and C_2 are given explicitly in (29).

Assumption 4. *There exist* $\dot{A} : \mathbb{R}^m \to \mathbb{R}^{d\times d}$, $\dot{b} > 0$ *such that for any* $x, y \in \mathbb{R}^d$,

$$\langle y, \dot{A}(x)y\rangle \geq 0$$

and for each $\theta, \theta' \in \mathbb{R}^d$ *satisfying* $|\theta - \theta'| > \dot{b}$, $x \in \mathbb{R}^m$,

$$\langle H(\theta, x) - H(\theta', x), \theta - \theta'\rangle \geq \langle \theta - \theta', \dot{A}(x)(\theta - \theta')\rangle. \tag{10}$$

The smallest eigenvalue of $\mathbb{E}[\dot{A}(X_0)]$ *is a positive real number* $\dot{a} > 0$.

Remark 2.4. *By Assumption 4 , one obtains for each* $\theta, \theta' \in \mathbb{R}^d$ *satisfying* $|\theta - \theta'| > \dot{b}$,

$$\left\langle h(\theta) - h(\theta'), \theta - \theta' \right\rangle \geq \dot{a}|\theta - \theta'|^2,\ \theta \in \mathbb{R}^d.$$

Theorem 2.5. *Let Assumptions 1, 2, and 4 hold. Then, there exist constants* $C_3, C_4, C_5 > 0$ *such that, for every* $0 < \lambda \leq \lambda_{\max}$, $\beta > 0$, *and* $n \in \mathbb{N}$,

$$\tilde{W}_1(\mathcal{L}(\theta^\lambda_n), \pi_\beta) \leq C_4 \mathrm{e}^{-C_3\lambda n}\mathbb{E}[|\theta_0|^4 + 1] + C_5\sqrt{\lambda}. \tag{11}$$

2.1 Related work and discussions

In Theorem 2.3 and 2.5, convergence results are provided in W_1 and $\tilde{W}_1$ distances with rate 1/2. Similar results are obtained in [1, Theorem 3.10] in the convex setting. Moreover, the analysis follows the approach in [5], nevertheless, we crucially extend its framework by assuming local Lipschitness of H in Assumption 2, and non-uniform estimates with respect to the x variable in Assumption 3 and 4.

Next, we mainly focus on the comparison of our work with [15] and [14]. In [15, Proposition 3.3], a finite-time convergence result of the SGLD algorithm in Wasserstein-2 distance is provided. To obtain this result, a dissipativity condition ([15, Assumption (A.3)]) is proposed. In [15, Assumption (A.1), (A.4)], the variance of the stochastic gradient is assumed to be bounded, as well as the quantities $f(0, \cdot)$ and $H(0, \cdot)$, where $U(\theta) = \mathbb{E}[f(\theta, X_0)]$ and $H(\cdot, \cdot) = \nabla_\theta f(\cdot, \cdot)$, $\theta \in \mathbb{R}^d$. In addition, it requires the finiteness of an exponential moment of the initial value ([15, Assumption (A.5)]) and the Lipschitz continuity of H in θ ([15, Assumption (A.2)]). In Theorem 2.3, we obtain a non-asymptotic result in W_1 distance (defined in (4)) with rate 1/2 under Assumptions 1, 2 and 3. While we improve the convergence rate $\lambda^{5/4}n$ in [15], our result is obtained in W_1 distance instead of $\tilde{W}_2$, and we further require a local Lipschitz continuity of $H(\theta, x)$ in x. However, compared to [15, Assumption (A.3)], we allow the dissipativity condition without imposing the uniformity in x in Assumption 3, and we require only polynomial moments of the initial value and the process $(X_n)_{n\in\mathbb{N}}$ in Assumption 1. Furthermore, in Assumption 2, we relax the Lipschitz condition of H in θ by allowing the Lipschitz constant to depend x. One notices that [15, Assumption (A.4)] can be obtained by using Assumption 1 and 2.

Compared to [14, Theorem 1.4] with $\alpha = 1$, Theorem 2.5 achieves the same rate in $\tilde{W}_1$ without assuming that the variance of the stochastic gradient is controlled by the stepsize ([14, Assumption 1.3]). Instead, we assume a local Lipschitz continuity of H in Assumption 2. Furthermore, the authors in [14] assume convexity at infinity of h whereas our Assumption 4 imposes the same condition for H but with non-uniform dependence in x.

3 Proofs

3.1 Further notation and introduction of auxiliary processses

Define the Lyapunov function for each $p \geq 1$ by

$$V_p(\theta) := (1 + |\theta|^2)^{p/2},\ \theta \in \mathbb{R}^d,$$

and similarly $\mathrm{v}_p(x) := (1+x^2)^{p/2}$, for any real $x \geq 0$. Notice that these functions are twice continuously differentiable and

$$\lim_{|\theta|\to\infty} \frac{\nabla V_p(\theta)}{V_p(\theta)} = 0.$$

Let $\mathcal{P}_{V_p}$ denote the set of $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfying $\int_{\mathbb{R}^d} V_p(\theta)\,\mu(\mathrm{d}\theta) < \infty$.

We define and discuss a number of auxiliary continuous-time processes below. First, for each $\lambda > 0$,

$$Z_t^\lambda := Z_{\lambda t},\ t \in \mathbb{R}_+.$$

Notice that $\tilde{B}_t^\lambda := B_{\lambda t}/\sqrt{\lambda}$, $t \in \mathbb{R}_+$ is also a Brownian motion and

$$\mathrm{d}Z_t^\lambda = -\lambda h(Z_t^\lambda)\,\mathrm{d}t + \sqrt{\frac{2\lambda}{\beta}}\mathrm{d}\tilde{B}_t^\lambda,\ Z_0^\lambda = \theta_0.$$

Denote by $\mathcal{F}_t$ the natural filtration of B_t, $t \in \mathbb{R}_+$. Then, $\mathcal{F}_t^\lambda := \mathcal{F}_{\lambda t}$, $t \in \mathbb{R}_+$ is the natural filtration of $\tilde{B}_t^\lambda$, $t \in \mathbb{R}_+$. One notice that $\mathcal{F}_t^\lambda$ is independent of $\mathcal{G}_\infty \vee \sigma(\theta_0)$. Then, define the continuous-time interpolation of the SGLD algorithm (2) as

$$\mathrm{d}\bar{\theta}_t^\lambda = -\lambda H(\bar{\theta}_{\lfloor t\rfloor}^\lambda, X_{\lceil t\rceil})\,\mathrm{d}t + \sqrt{\frac{2\lambda}{\beta}}\mathrm{d}\tilde{B}_t^\lambda, \tag{12}$$

with initial condition $\bar{\theta}_0^\lambda = \theta_0$. In addition, for each integer $n \in \mathbb{N}$,

$$\mathcal{L}(\bar{\theta}_n^\lambda) = \mathcal{L}(\theta_n^\lambda). \tag{13}$$

Due to the homogeneous nature of the coefficients of equation (12), the law of the interpolated process coincides with the law of the SGLD algorithm (2) at grid-points. Hence, crucial estimates for the SGLD can be derived by studying equation (12).

Furthermore, consider a continuous-time process $\zeta_t^{s,v,\lambda}$, $t \geq s$, which denotes the solution of the SDE

$$\mathrm{d}\zeta_t^{s,v,\lambda} = -\lambda h(\zeta_t^{s,v,\lambda})\mathrm{d}t + \sqrt{\frac{2\lambda}{\beta}}\mathrm{d}\tilde{B}_t^\lambda.$$

with initial condition $\zeta_s^{s,v,\lambda} := v$, $v \in \mathbb{R}^d$.

Definition 3.1. *Fix $n \in \mathbb{N}$ and define*

$$\bar{\zeta}_t^{\lambda,n} = \zeta_t^{nT,\bar{\theta}_{nT}^\lambda,\lambda}$$

where $T := \lfloor 1/\lambda \rfloor$.

Intuitively, $\bar{\zeta}_t^{\lambda,n}$ is a process started from the value of the SGLD process (12) at time nT and made run until time $t \geq nT$ with the continuous-time Langevin dynamics.

3.2 Preliminary estimates

It is a classic result that SDE (1) has a unique solution adapted to $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$, since h is Lipschitz-continuous by (6). In order to obtain the convergence results, we first establish the moment bounds of the process $(\bar{\theta}_t^\lambda)_{t\geq 0}$.

Lemma 3.2. *Let Assumptions 1, 2 and 3 hold. For any $0 < \lambda < \lambda_{\max}$ given in (8), $n \in \mathbb{N}$, $t \in (n, n+1]$,*

$$\mathbb{E}|\bar{\theta}_t^\lambda|^2 \leq (1 - a\lambda(t-n))(1-a\lambda)^n \mathbb{E}|\theta_0|^2 + c_1(\lambda_{\max} + a^{-1}),$$

where

$$c_1 = (c_0 + 2d/\beta), \quad c_0 = 4\lambda_{\max} L_2^2 \mathbb{E}\left[(1+|X_0|)^{2\rho+2}\right] + 4\lambda_{\max} H_\star^2 + 2b. \tag{14}$$

In addition, $\sup_t \mathbb{E}|\bar{\theta}_t^\lambda|^2 \leq \mathbb{E}|\theta_0|^2 + c_1(\lambda_{\max} + a^{-1}) < \infty$. Similarly, one obtains

$$\mathbb{E}\left[|\bar{\theta}_t^\lambda|^4\right] \leq (1 - a\lambda(t-n))(1-a\lambda)^n \mathbb{E}|\theta_0|^4 + c_3(\lambda_{\max} + a^{-1}),$$

where

$$c_3 = (1 + a\lambda_{\max})c_2 + 12d^2\beta^{-2}(\lambda_{\max} + 9a^{-1}) \tag{15}$$

with c_2 given in (20). Moreover, this implies $\sup_t \mathbb{E}|\bar{\theta}_t^\lambda|^4 < \infty$.

Proof. For any $n \in \mathbb{N}$ and $t \in (n, n+1]$, define $\Delta_{n,t} = \bar{\theta}^{\lambda}_n - \lambda H(\bar{\theta}^{\lambda}_n, X_{n+1})(t-n)$. By using (12), it is easily seen that for $t \in (n, n+1]$

$$\mathbb{E}\left[|\bar{\theta}^{\lambda}_t|^2 \,\middle|\, \bar{\theta}^{\lambda}_n\right] = \mathbb{E}\left[|\Delta_{n,t}|^2 \,\middle|\, \bar{\theta}^{\lambda}_n\right] + (2\lambda/\beta)d(t-n).$$

Then, by using Assumptions 1, 2, Remark 2.1 and 3, one obtains

$$\begin{aligned}
&\mathbb{E}\left[|\Delta_{n,t}|^2 \,\middle|\, \bar{\theta}^{\lambda}_n\right] \\
&= |\bar{\theta}^{\lambda}_n|^2 - 2\lambda(t-n)\mathbb{E}\left[\left\langle \bar{\theta}^{\lambda}_n, H(\bar{\theta}^{\lambda}_n, X_{n+1})\right\rangle \,\middle|\, \bar{\theta}^{\lambda}_n\right] \\
&\quad + \lambda^2(t-n)^2\mathbb{E}\left[|H(\bar{\theta}^{\lambda}_n, X_{n+1})|^2 \,\middle|\, \bar{\theta}^{\lambda}_n\right] \\
&\le |\bar{\theta}^{\lambda}_n|^2 - 2\lambda(t-n)\left\langle \bar{\theta}^{\lambda}_n, \mathbb{E}\left[A(X_0)\right]\bar{\theta}^{\lambda}_n\right\rangle + 2\lambda(t-n)b \\
&\quad + \lambda^2(t-n)^2\mathbb{E}\left[|L_1(1+|X_{n+1}|)^{\rho}|\bar{\theta}^{\lambda}_n| + L_2(1+|X_{n+1}|)^{\rho+1} + H_{\star}|^2 \,\middle|\, \bar{\theta}^{\lambda}_n\right] \\
&\le (1-2a\lambda(t-n))|\bar{\theta}^{\lambda}_n|^2 + 2\lambda^2(t-n)^2L_1^2\mathbb{E}\left[(1+|X_0|)^{2\rho}\right]|\bar{\theta}^{\lambda}_n|^2 \\
&\quad + 4\lambda^2(t-n)^2L_2^2\mathbb{E}\left[(1+|X_0|)^{2\rho+2}\right] + 4\lambda^2(t-n)^2H_{\star}^2 + 2\lambda(t-n)b,
\end{aligned}$$

where the last inequality is obtained by using $(a+b)^2 \le 2a^2 + 2b^2$, for $a, b \ge 0$ twice. For $\lambda < \lambda_{\max}$,

$$\mathbb{E}\left[|\Delta_{n,t}|^2 \,\middle|\, \bar{\theta}^{\lambda}_n\right] \le (1-a\lambda(t-n))|\bar{\theta}^{\lambda}_n|^2 + \lambda(t-n)c_0,$$

where $c_0 = 4\lambda_{\max}L_2^2\mathbb{E}\left[(1+|X_0|)^{2\rho+2}\right] + 4\lambda_{\max}H_{\star}^2 + 2b$. Therefore, one obtains

$$\mathbb{E}\left[|\bar{\theta}^{\lambda}_t|^2 \,\middle|\, \bar{\theta}^{\lambda}_n\right] \le (1-a\lambda(t-n))|\bar{\theta}^{\lambda}_n|^2 + \lambda(t-n)c_1,$$

where $c_1 = (c_0 + 2d/\beta)$ and the result follows by induction. To calculate a higher moment, denote by $U^{\lambda}_{n,t} = \{2\lambda\beta^{-1}\}^{1/2}(\tilde{B}^{\lambda}_t - \tilde{B}^{\lambda}_n)$, for $t \in (n, n+1]$, one calculates

$$\begin{aligned}
\mathbb{E}\left[|\bar{\theta}^{\lambda}_t|^4 \,\middle|\, \bar{\theta}^{\lambda}_n\right] &= \mathbb{E}\left[\left(|\Delta_{n,t}|^2 + |U^{\lambda}_{n,t}|^2 + 2\left\langle \Delta_{n,t}, U^{\lambda}_{n,t}\right\rangle\right)^2 \,\middle|\, \bar{\theta}^{\lambda}_n\right] \\
&= \mathbb{E}\Big[|\Delta_{n,t}|^4 + |U^{\lambda}_{n,t}|^4 + 2|\Delta_{n,t}|^2|U^{\lambda}_{n,t}|^2 + 4|\Delta_{n,t}|^2\left\langle \Delta_{n,t}, U^{\lambda}_{n,t}\right\rangle \\
&\qquad + 4|U^{\lambda}_{n,t}|^2\left\langle \Delta_{n,t}, U^{\lambda}_{n,t}\right\rangle + 4\left(\left\langle \Delta_{n,t}, U^{\lambda}_{n,t}\right\rangle\right)^2 \,\Big|\, \bar{\theta}^{\lambda}_n\Big] \\
&\le \mathbb{E}\left[|\Delta_{n,t}|^4 + |U^{\lambda}_{n,t}|^4 + 6|\Delta_{n,t}|^2|U^{\lambda}_{n,t}|^2 \,\middle|\, \bar{\theta}^{\lambda}_n\right] \\
&\le (1+a\lambda(t-n))\mathbb{E}\left[|\Delta_{n,t}|^4 \,\middle|\, \bar{\theta}^{\lambda}_n\right] + (1+9/(a\lambda(t-n)))\mathbb{E}\left[|U^{\lambda}_{n,t}|^4\right].
\end{aligned} \tag{16}$$

where the last inequality holds due to $2ab \le \epsilon a^2 + \epsilon^{-1}b^2$, for $a, b \ge 0$ and $\epsilon > 0$. Then, one continues with calculating

$$\begin{aligned}
\mathbb{E}\left[|\Delta_{n,t}|^4 \,\middle|\, \bar{\theta}^{\lambda}_n\right] &= \mathbb{E}\left[\left(|\bar{\theta}^{\lambda}_n|^2 - 2\lambda(t-n)\left\langle \bar{\theta}^{\lambda}_n, H(\bar{\theta}^{\lambda}_n, X_{n+1})\right\rangle + \lambda^2(t-n)^2|H(\bar{\theta}^{\lambda}_n, X_{n+1})|^2\right)^2 \,\middle|\, \bar{\theta}^{\lambda}_n\right] \\
&\le |\bar{\theta}^{\lambda}_n|^4 + \mathbb{E}\Big[6\lambda^2(t-n)^2|\bar{\theta}^{\lambda}_n|^2|H(\bar{\theta}^{\lambda}_n, X_{n+1})|^2 - 4\lambda(t-n)\left\langle \bar{\theta}^{\lambda}_n, H(\bar{\theta}^{\lambda}_n, X_{n+1})\right\rangle|\bar{\theta}^{\lambda}_n|^2 \\
&\qquad - 4\lambda^3(t-n)^3|H(\bar{\theta}^{\lambda}_n, X_{n+1})|^2\left\langle \bar{\theta}^{\lambda}_n, H(\bar{\theta}^{\lambda}_n, X_{n+1})\right\rangle + \lambda^4(t-n)^4|H(\bar{\theta}^{\lambda}_n, X_{n+1})|^4 \,\Big|\, \bar{\theta}^{\lambda}_n\Big].
\end{aligned}$$

Observing that, by Remark 2.1, for $q \ge 1$,

$$\mathbb{E}\left[|H(\bar{\theta}^{\lambda}_n, X_{n+1})|^q \,\middle|\, \bar{\theta}^{\lambda}_n\right] \le 2^{q-1}L_1^q\mathbb{E}\left[(1+|X_0|)^{q\rho}\right]|\bar{\theta}^{\lambda}_n|^q + 2^{2q-2}L_2^q\mathbb{E}\left[|1+|X_0||^{q\rho+q}\right] + 2^{2q-2}H_{\star}^q. \tag{17}$$

By using Assumption 3 and by taking $q = 2, 3, 4$ in (17), one obtains

$$\mathbb{E}\left[|\Delta_{n,t}|^4 \,\middle|\, \bar{\theta}^{\lambda}_n\right] \le (1-4a\lambda(t-n))|\bar{\theta}^{\lambda}_n|^4 + 4b\lambda(t-n)|\bar{\theta}^{\lambda}_n|^2$$

$$
\begin{aligned}
&+ 12\lambda^2(t-n)^2 L_1^2 \mathbb{E}\left[(1+|X_0|)^{2\rho}\right] |\bar{\theta}_n^\lambda|^4 + 16\lambda^3(t-n)^3 L_1^3 \mathbb{E}\left[(1+|X_0|)^{3\rho}\right] |\bar{\theta}_n^\lambda|^4 \\
&+ 8\lambda^4(t-n)^4 L_1^4 \mathbb{E}\left[(1+|X_0|)^{4\rho}\right] |\bar{\theta}_n^\lambda|^4 \\
&+ 24\lambda^2(t-n)^2 \left(L_2^2 \mathbb{E}\left[(1+|X_0|)^{2\rho+2}\right] + H_\star^2\right) |\bar{\theta}_n^\lambda|^2 \\
&+ 64\lambda^3(t-n)^3 \left(L_2^3 \mathbb{E}\left[(1+|X_0|)^{3\rho+3}\right] + H_\star^3\right) |\bar{\theta}_n^\lambda| \\
&+ 64\lambda^4(t-n)^4 \left(L_2^4 \mathbb{E}\left[(1+|X_0|)^{4\rho+4}\right] + H_\star^4\right)
\end{aligned}
$$

which implies, by using $\lambda < \lambda_{\max}$

$$
\begin{aligned}
\mathbb{E}\left[|\Delta_{n,t}|^4 \,\middle|\, \bar{\theta}_n^\lambda\right] \leq (1 - 3a\lambda(t-n))|\bar{\theta}_n^\lambda|^4 &+ 4b\lambda(t-n)|\bar{\theta}_n^\lambda|^2 \\
&+ 24\lambda^2(t-n)^2 \left(L_2^2 \mathbb{E}\left[(1+|X_0|)^{2\rho+2}\right] + H_\star^2\right) |\bar{\theta}_n^\lambda|^2 \\
&+ 64\lambda^3(t-n)^3 \left(L_2^3 \mathbb{E}\left[(1+|X_0|)^{3\rho+3}\right] + H_\star^3\right) |\bar{\theta}_n^\lambda| \\
&+ 64\lambda^4(t-n)^4 \left(L_2^4 \mathbb{E}\left[(1+|X_0|)^{4\rho+4}\right] + H_\star^4\right).
\end{aligned}
$$

For $|\bar{\theta}_n^\lambda| > (8ba^{-1} + 48a^{-1}\lambda_{\max}(L_2^2\mathbb{E}\left[(1+|X_0|)^{2\rho+2}\right] + H_\star^2))^{1/2}$, we have

$$
-\frac{a\lambda(t-n)}{2}|\bar{\theta}_n^\lambda|^4 + 4b\lambda(t-n)|\bar{\theta}_n^\lambda|^2 + 24\lambda^2(t-n)^2 \left(L_2^2 \mathbb{E}\left[(1+|X_0|)^{2\rho+2}\right] + H_\star^2\right) |\bar{\theta}_n^\lambda|^2 < 0,
$$

similarly, for $|\bar{\theta}_n^\lambda| > (128a^{-1}\lambda_{\max}^2(L_2^3\mathbb{E}\left[(1+|X_0|)^{3\rho+3}\right] + H_\star^3))^{1/3}$

$$
-\frac{a\lambda(t-n)}{2}|\bar{\theta}_n^\lambda|^4 + 64\lambda^3(t-n)^3 \left(L_2^3 \mathbb{E}\left[(1+|X_0|)^{3\rho+3}\right] + H_\star^3\right) |\bar{\theta}_n^\lambda| < 0.
$$

Denote by

$$
\begin{aligned}
M = \max\{&(8ba^{-1} + 48a^{-1}\lambda_{\max}(L_2^2\mathbb{E}\left[(1+|X_0|)^{2\rho+2}\right] + H_\star^2))^{1/2}, \\
&(128a^{-1}\lambda_{\max}^2(L_2^3\mathbb{E}\left[(1+|X_0|)^{3\rho+3}\right] + H_\star^3))^{1/3}\}.
\end{aligned} \tag{18}
$$

For $|\bar{\theta}_n^\lambda| > M$, one obtains

$$
\mathbb{E}\left[|\Delta_{n,t}|^4 \,\middle|\, \bar{\theta}_n^\lambda\right] \leq (1 - 2a\lambda(t-n))|\bar{\theta}_n^\lambda|^4 + 64\lambda^4(t-n)^4 \left(L_2^4 \mathbb{E}\left[(1+|X_0|)^{4\rho+4}\right] + H_\star^4\right).
$$

As for $|\bar{\theta}_n^\lambda| \leq M$, we have

$$
\begin{aligned}
\mathbb{E}\left[|\Delta_{n,t}|^4 \,\middle|\, \bar{\theta}_n^\lambda\right] \leq (1 - 2a\lambda(t-n))|\bar{\theta}_n^\lambda|^4 &+ 4b\lambda(t-n)M^2 \\
&+ 24\lambda^2(t-n)^2 \left(L_2^2 \mathbb{E}\left[(1+|X_0|)^{2\rho+2}\right] + H_\star^2\right) M^2 \\
&+ 64\lambda^3(t-n)^3 \left(L_2^3 \mathbb{E}\left[(1+|X_0|)^{3\rho+3}\right] + H_\star^3\right) M \\
&+ 64\lambda^4(t-n)^4 \left(L_2^4 \mathbb{E}\left[(1+|X_0|)^{4\rho+4}\right] + H_\star^4\right).
\end{aligned}
$$

Combining the two cases yields

$$
\mathbb{E}\left[|\Delta_{n,t}|^4 \,\middle|\, \bar{\theta}_n^\lambda\right] \leq (1 - 2a\lambda(t-n))|\bar{\theta}_n^\lambda|^4 + \lambda(t-n)c_2, \tag{19}
$$

where

$$
c_2 = 4bM^2 + 152(1+\lambda_{\max})^3 \left((1+L_2)^4\mathbb{E}\left[(1+|X_0|)^{4\rho+4}\right] + (1+H_\star)^4\right)(1+M)^2 \tag{20}
$$

with M given in (18). Substituting (19) into (16), one obtains

$$
\begin{aligned}
\mathbb{E}\left[|\bar{\theta}_t^\lambda|^4 \,\middle|\, \bar{\theta}_n^\lambda\right] &\leq (1 + a\lambda(t-n))(1 - 2a\lambda(t-n))|\bar{\theta}_n^\lambda|^4 \\
&\quad + (1 + a\lambda(t-n))\lambda(t-n)c_2 + 12d^2\lambda^2\beta^{-2}(t-n)^2(1 + 9/(a\lambda(t-n))) \\
&\leq (1 - a\lambda(t-n))|\bar{\theta}_n^\lambda|^4 + \lambda(t-n)c_3,
\end{aligned}
$$

where $c_3 = (1 + a\lambda_{\max})c_2 + 12d^2\beta^{-2}(\lambda_{\max} + 9a^{-1})$. The proof completes by induction. $\square$

Corollary 3.3. *Let Assumptions 1, 2 and 3 hold. For any* $0<\lambda<\lambda_{\max}$ *given in* (8), $n\in\mathbb{N}$, $t\in(n,n+1]$,

$$\mathbb{E}[V_4(\bar{\theta}_t^\lambda)]\leq 2(1-a\lambda)^{\lfloor t\rfloor}\mathbb{E}[V_4(\theta_0)]+2c_3(\lambda_{\max}+a^{-1})+2,$$

where c_3 *is given in* (15).

Next, we present a drift condition associated with the SDE (1), which will be used to obtain the moment bounds of the process $\bar{\zeta}_t^{\lambda,n}$.

Lemma 3.4. *Let Assumption 3 holds. Then, for each* $p\geq 2$, $\theta\in\mathbb{R}^d$,

$$\frac{\Delta V_p}{\beta}-\langle h(\theta),\nabla V_p(\theta)\rangle\leq -\bar{c}(p)V_p(\theta)+\tilde{c}(p),$$

where $\bar{c}(p)=ap/4$ *and* $\tilde{c}(p)=(3/4)ap\mathrm{v}_p(\overline{M}_p)$ *with* $\overline{M}_p=\sqrt{1/3+4b/(3a)+4d/(3a\beta)+4(p-2)/(3a\beta)}$.

Proof. See [5, Lemma 3.6]. □

Lemma 3.5. *Let Assumptions 1, 2 and 3 hold. For any* $0<\lambda<\lambda_{\max}$ *given in* (8), $t\geq nT$, $n\in\mathbb{N}$, *one obtains the following inequality*

$$\mathbb{E}[V_2(\bar{\zeta}_t^{\lambda,n})]\leq e^{-a\lambda t/2}\mathbb{E}[V_2(\theta_0)]+3\mathrm{v}_2(\bar{M}_2)+c_1(\lambda_{\max}+a^{-1})+1,$$

where the process $\bar{\zeta}_t^{\lambda,n}$ *is defined in Definition 3.1 and* c_1 *is given in* (14). *Furthermore,*

$$\mathbb{E}[V_4(\bar{\zeta}_t^{\lambda,n})]\leq 2e^{-a\lambda t}\mathbb{E}[V_4(\theta_0)]+3\mathrm{v}_4(\bar{M}_4)+2c_3(\lambda_{\max}+a^{-1})+2,$$

where c_3 *is given in* (15).

Proof. For any $p\geq 1$, application of Ito's lemma and taking expectation yields

$$\mathbb{E}[V_p(\bar{\zeta}_t^{\lambda,n})]=\mathbb{E}[V_p(\bar{\theta}_{nT}^\lambda)]+\int_{nT}^t\mathbb{E}\left[\lambda\frac{\Delta V_p(\bar{\zeta}_s^{\lambda,n})}{\beta}-\lambda\langle h(\bar{\zeta}_s^{\lambda,n}),\nabla V_p(\bar{\zeta}_s^{\lambda,n})\rangle\right]\mathrm{d}s.$$

Differentiating both sides and using Lemma 3.4, we arrive at

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[V_p(\bar{\zeta}_t^{\lambda,n})]=\mathbb{E}\left[\lambda\frac{\Delta V_p(\bar{\zeta}_t^{\lambda,n})}{\beta}-\lambda\langle h(\bar{\zeta}_t^{\lambda,n}),\nabla V_p(\bar{\zeta}_t^{\lambda,n})\rangle\right]\leq-\lambda\bar{c}(p)\mathbb{E}[V_p(\bar{\zeta}_t^{\lambda,n})]+\lambda\tilde{c}(p),$$

which yields

$$\begin{aligned}\mathbb{E}[V_p(\bar{\zeta}_t^{\lambda,n})]&\leq e^{-\lambda(t-nT)\bar{c}(p)}\mathbb{E}[V_p(\bar{\theta}_{nT}^\lambda)]+\frac{\tilde{c}(p)}{\bar{c}(p)}\left(1-e^{-\lambda\bar{c}(p)(t-nT)}\right)\\&\leq e^{-\lambda(t-nT)\bar{c}(p)}\mathbb{E}[V_p(\bar{\theta}_{nT}^\lambda)]+\frac{\tilde{c}(p)}{\bar{c}(p)}.\end{aligned}$$

Now for $p=2$, using Corollary 3.3 and Lemma 3.4, we obtain

$$\begin{aligned}\mathbb{E}[V_2(\bar{\zeta}_t^{\lambda,n})]&\leq e^{-\lambda(t-nT)\bar{c}(2)}\mathbb{E}[V_2(\bar{\theta}_{nT}^\lambda)]+\frac{\tilde{c}(2)}{\bar{c}(2)}\\&\leq(1-a\lambda)^{nT}e^{-\lambda(t-nT)\bar{c}(2)}\mathbb{E}[V_2(\theta_0)]+\frac{\tilde{c}(2)}{\bar{c}(2)}+c_1(\lambda_{\max}+a^{-1})+1\\&\leq e^{-a\lambda t/2}\mathbb{E}[V_2(\theta_0)]+3\mathrm{v}_2(\bar{M}_2)+c_1(\lambda_{\max}+a^{-1})+1,\end{aligned}$$

where the last inequality holds due to $1-z\leq e^{-z}$ for $z\geq 0$ and $\bar{c}(2)=a/2$. Similarly, for $p=4$, one obtains

$$\begin{aligned}\mathbb{E}[V_4(\bar{\zeta}_t^{\lambda,n})]&\leq e^{-\lambda(t-nT)\bar{c}(4)}\mathbb{E}[V_4(\bar{\theta}_{nT}^\lambda)]+\frac{\tilde{c}(4)}{\bar{c}(4)}\\&\leq 2(1-a\lambda)^{nT}e^{-\lambda(t-nT)\bar{c}(4)}\mathbb{E}[V_4(\theta_0)]+\frac{\tilde{c}(4)}{\bar{c}(4)}+2c_3(\lambda_{\max}+a^{-1})+2\\&\leq 2e^{-a\lambda t}\mathbb{E}[V_4(\theta_0)]+3\mathrm{v}_4(\bar{M}_4)+2c_3(\lambda_{\max}+a^{-1})+2,\end{aligned}$$

where the last inequality holds due to $1-z\leq e^{-z}$ for $z\geq 0$ and $\bar{c}(4)=a$. □

3.3 Proof of the main theorems

We introduce a functional which is crucial to obtain the convergence rate in W_1. For any $p \geq 1$, $\mu, \nu \in \mathcal{P}_{V_p}$,

$$w_{1,p}(\mu,\nu) := \inf_{\zeta \in \mathcal{C}(\mu,\nu)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} [1 \wedge |\theta - \theta'|](1 + V_p(\theta) + V_p(\theta'))\zeta(\mathrm{d}\theta \mathrm{d}\theta'), \tag{21}$$

and it satisfies trivially

$$W_1(\mu,\nu) \leq w_{1,p}(\mu,\nu). \tag{22}$$

The case $p = 2$, i.e. $w_{1,2}$, is used throughout the section. The result below states a contraction property of $w_{1,2}$.

Proposition 3.6. *Let Z'_t, $t \in \mathbb{R}_+$ be the solution of (1) with initial condition $Z'_0 = \theta_0$ which is independent of $\mathcal{F}_\infty$ and satisfies $|\theta_0| \in L^2$. Then,*

$$w_{1,2}(\mathcal{L}(Z_t), \mathcal{L}(Z'_t)) \leq \hat{c} e^{-\dot{c}t} w_{1,2}(\mathcal{L}(\theta_0), \mathcal{L}(\theta'_0)),$$

where the constants $\dot{c}$ and $\hat{c}$ are given in Lemma 3.10.

Proof. See Proposition 3.2 of [5]. □

In the following lemmas, we construct the non-asymptotic bound between $\bar{\theta}^\lambda_t$ and Z^λ_t, $t \in [nT, (n+1)T]$, in W_1 distance by decomposing the error using the auxiliary process $\bar{\zeta}^{\lambda,n}_t$:

$$W_1(\mathcal{L}(\bar{\theta}^\lambda_t), \mathcal{L}(Z^\lambda_t)) \leq W_1(\mathcal{L}(\bar{\theta}^\lambda_t), \mathcal{L}(\bar{\zeta}^{\lambda,n}_t)) + W_1(\mathcal{L}(\bar{\zeta}^{\lambda,n}_t), \mathcal{L}(Z^\lambda_t)). \tag{23}$$

One notices that when $1 < \lambda \leq \lambda_{\max}$, the result holds trivially. Thus, we consider the case $0 < \lambda \leq 1$, which implies $1/2 < \lambda T \leq 1$.

Lemma 3.7. *Let Assumption 1, 2 and 3 hold. For any $0 < \lambda < \lambda_{\max}$ given in (8), $t \in [nT, (n+1)T]$,*

$$W_1(\mathcal{L}(\bar{\theta}^\lambda_t), \mathcal{L}(\bar{\zeta}^{\lambda,n}_t)) \leq \tilde{W}_2(\mathcal{L}(\bar{\theta}^\lambda_t), \mathcal{L}(\bar{\zeta}^{\lambda,n}_t)) \leq \sqrt{\lambda}(e^{-an/4}\bar{C}_{2,1}\mathbb{E}[V_2(\theta_0)] + \bar{C}_{2,2})^{1/2},$$

where $\bar{C}_{2,1}$ and $\bar{C}_{2,2}$ are given in (26).

Proof. To handle the first term in (23), we start by establishing an upper bound in Wasserstein-2 distance and the statment follows by noticing $W_1 \leq \tilde{W}_2$. By employing synchronous coupling, using (12) and the definition of $\bar{\zeta}^{\lambda,n}_t$ in Definition 3.1, one obtains

$$\left|\bar{\zeta}^{\lambda,n}_t - \bar{\theta}^\lambda_t\right| \leq \lambda \left| \int_{nT}^t \left[H(\bar{\theta}^\lambda_{\lfloor s \rfloor}, X_{\lceil s \rceil}) - h(\bar{\zeta}^{\lambda,n}_s)\right] \mathrm{d}s \right|.$$

Then, the triangle inequality leads

$$\left|\bar{\zeta}^{\lambda,n}_t - \bar{\theta}^\lambda_t\right| \leq \lambda \left| \int_{nT}^t \left[H(\bar{\theta}^\lambda_{\lfloor s \rfloor}, X_{\lceil s \rceil}) - H(\bar{\zeta}^{\lambda,n}_s, X_{\lceil s \rceil})\right] \mathrm{d}s \right| + \lambda \left| \int_{nT}^t \left[h(\bar{\zeta}^{\lambda,n}_s) - H(\bar{\zeta}^{\lambda,n}_s, X_{\lceil s \rceil})\right] \mathrm{d}s \right|.$$

Applying Assumption 2, we obtain

$$\left|\bar{\zeta}^{\lambda,n}_t - \bar{\theta}^\lambda_t\right| \leq \lambda L_1 \int_{nT}^t \left(1 + |X_{\lceil s \rceil}|\right)^\rho \left|\bar{\theta}^\lambda_{\lfloor s \rfloor} - \bar{\zeta}^{\lambda,n}_s\right| \mathrm{d}s + \lambda \left| \int_{nT}^t \left[h(\bar{\zeta}^{\lambda,n}_s) - H(\bar{\zeta}^{\lambda,n}_s, X_{\lceil s \rceil})\right] \mathrm{d}s \right|$$

Now taking squares of both sides and using $(a+b)^2 \leq 2a^2 + 2b^2$ for $a, b > 0$ lead to

$$\left|\bar{\zeta}^{\lambda,n}_t - \bar{\theta}^\lambda_t\right|^2 \leq 2\lambda L_1^2 \int_{nT}^t (1 + |X_{\lceil s \rceil}|)^{2\rho} \left|\bar{\theta}^\lambda_{\lfloor s \rfloor} - \bar{\zeta}^{\lambda,n}_s\right|^2 \mathrm{d}s + 2\lambda^2 \left| \int_{nT}^t \left[h(\bar{\zeta}^{\lambda,n}_s) - H(\bar{\zeta}^{\lambda,n}_s, X_{\lceil s \rceil})\right] \mathrm{d}s \right|^2. \tag{24}$$

Taking expectations yields

$$
\begin{aligned}
&\mathbb{E}\left[\left|\bar{\zeta}_t^{\lambda,n}-\bar{\theta}_t^{\lambda}\right|^2\right]\\
&\leq 2\lambda L_1^2\int_{nT}^t \mathbb{E}\left[(1+|X_{\lceil s\rceil}|)^{2\rho}\right]\mathbb{E}\left[\left|\bar{\theta}_{\lfloor s\rfloor}^{\lambda}-\bar{\zeta}_s^{\lambda,n}\right|^2\right]\mathrm{d}s+2\lambda^2\mathbb{E}\left[\left|\int_{nT}^t\left[h(\bar{\zeta}_s^{\lambda,n})-H(\bar{\zeta}_s^{\lambda,n},X_{\lceil s\rceil})\right]\mathrm{d}s\right|^2\right],\\
&\leq 2\lambda L_1^2 C_\rho\int_{nT}^t \mathbb{E}\left[\left|\bar{\theta}_{\lfloor s\rfloor}^{\lambda}-\bar{\zeta}_s^{\lambda,n}\right|^2\right]\mathrm{d}s+2\lambda^2\mathbb{E}\left[\left|\int_{nT}^t\left[h(\bar{\zeta}_s^{\lambda,n})-H(\bar{\zeta}_s^{\lambda,n},X_{\lceil s\rceil})\right]\mathrm{d}s\right|^2\right],
\end{aligned}
$$

where C_ρ is defined in (7) and the expectation splits over terms in the first integral due to the independence of $X_{\lceil s\rceil}$ from the rest of the random variables. Using $\lambda T\leq 1$, Lemma 5.2 and $(a+b)^2\leq 2a^2+2b^2$ once again, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\left|\bar{\zeta}_t^{\lambda,n}-\bar{\theta}_t^{\lambda}\right|^2\right]&\leq 4\lambda L_1^2C_\rho\int_{nT}^t\mathbb{E}\left[\left|\bar{\theta}_{\lfloor s\rfloor}^{\lambda}-\bar{\theta}_s^{\lambda}\right|^2\right]\mathrm{d}s+4\lambda L_1^2C_\rho\int_{nT}^t\mathbb{E}\left[\left|\bar{\theta}_s^{\lambda}-\bar{\zeta}_s^{\lambda,n}\right|^2\right]\mathrm{d}s\\
&\quad+2\lambda^2\mathbb{E}\left[\left|\int_{nT}^t\left[h(\bar{\zeta}_s^{\lambda,n})-H(\bar{\zeta}_s^{\lambda,n},X_{\lceil s\rceil})\right]\mathrm{d}s\right|^2\right]\\
&\leq 4\lambda L_1^2C_\rho(e^{-a\lambda nT}\bar{\sigma}_Y\mathbb{E}[V_2(\theta_0)]+\tilde{\sigma}_Y)+4\lambda L_1^2C_\rho\int_{nT}^t\mathbb{E}\left[\left|\bar{\theta}_s^{\lambda}-\bar{\zeta}_s^{\lambda,n}\right|^2\right]\mathrm{d}s\\
&\quad+2\lambda^2\mathbb{E}\left[\left|\int_{nT}^t\left[h(\bar{\zeta}_s^{\lambda,n})-H(\bar{\zeta}_s^{\lambda,n},X_{\lceil s\rceil})\right]\mathrm{d}s\right|^2\right].
\end{aligned}
\tag{25}
$$

where $\bar{\sigma}_Y$ and $\tilde{\sigma}_Y$ are provided in (31). Next, we bound the last term by partitioning the last integral. Assume that $nT+K\leq t\leq nT+K+1$ where $K+1\leq T$. Thus we can write

$$
\left|\int_{nT}^t\left[h(\bar{\zeta}_s^{\lambda,n})-H(\bar{\zeta}_s^{\lambda,n},X_{\lceil s\rceil})\right]\mathrm{d}s\right|=\left|\sum_{k=1}^K I_k+R_K\right|
$$

where

$$
I_k=\int_{nT+(k-1)}^{nT+k}[h(\bar{\zeta}_s^{\lambda,n})-H(\bar{\zeta}_s^{\lambda,n},X_{nT+k})]\mathrm{d}s\qquad\text{and}\qquad R_K=\int_{nT+K}^t[h(\bar{\zeta}_s^{\lambda,n})-H(\bar{\zeta}_s^{\lambda,n},X_{nT+K+1})]\mathrm{d}s.
$$

Taking squares of both sides

$$
\left|\sum_{k=1}^K I_k+R_K\right|^2=\sum_{k=1}^K|I_k|^2+2\sum_{k=2}^K\sum_{j=1}^{k-1}\langle I_k,I_j\rangle+2\sum_{k=1}^K\langle I_k,R_K\rangle+|R_K|^2,
$$

Finally, we take expectations of both sides. Define the filtration $\mathcal{H}_t=\mathcal{F}_\infty^\lambda\vee\mathcal{G}_{\lfloor t\rfloor}$. We first note that for any $k=2,\dots,K$, $j=1,\dots,k-1$,

$$
\begin{aligned}
&\mathbb{E}\langle I_k,I_j\rangle\\
&=\mathbb{E}\left[\mathbb{E}[\langle I_k,I_j\rangle|\mathcal{H}_{nT+j}]\right],\\
&=\mathbb{E}\left[\mathbb{E}\left[\left\langle\int_{nT+(k-1)}^{nT+k}[H(\bar{\zeta}_s^{\lambda,n},X_{nT+k})-h(\bar{\zeta}_s^{\lambda,n})]\mathrm{d}s,\int_{nT+(j-1)}^{nT+j}[H(\bar{\zeta}_s^{\lambda,n},X_{nT+j})-h(\bar{\zeta}_s^{\lambda,n})]\mathrm{d}s\right\rangle\middle|\mathcal{H}_{nT+j}\right]\right],\\
&=\mathbb{E}\left[\left\langle\int_{nT+(k-1)}^{nT+k}\mathbb{E}\left[H(\bar{\zeta}_s^{\lambda,n},X_{nT+k})-h(\bar{\zeta}_s^{\lambda,n})\middle|\mathcal{H}_{nT+j}\right]\mathrm{d}s,\int_{nT+(j-1)}^{nT+j}[H(\bar{\zeta}_s^{\lambda,n},X_{nT+j})-h(\bar{\zeta}_s^{\lambda,n})]\mathrm{d}s\right\rangle\right],\\
&=0.
\end{aligned}
$$

By the same argument $\mathbb{E}\langle I_k, R_K\rangle = 0$ for all $1 \le k \le K$. Therefore, the last term of (25) is bounded as

$$2\lambda^2\mathbb{E}\left[\left|\int_{nT}^{t}\left[h(\bar{\zeta}_s^{\lambda,n}) - H(\bar{\zeta}_s^{\lambda,n}, X_{\lceil s\rceil})\right]\mathrm{d}s\right|^2\right] = 2\lambda^2\sum_{k=1}^{K}\mathbb{E}\left[|I_k|^2\right] + 2\lambda^2\mathbb{E}\left[|R_K|^2\right]$$
$$\le 4e^{-a\lambda nT/2}\lambda(\bar{\sigma}_Z\mathbb{E}[V_2(\theta_0)] + \tilde{\sigma}_Z),$$

where the last inequality holds due to Lemma 5.1 and $\bar{\sigma}_Z$ and $\tilde{\sigma}_Z$ are provided in (30). Therefore, the bound (25) becomes

$$\mathbb{E}\left[\left|\bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^{\lambda}\right|^2\right] \le 4\lambda L_1^2 C_\rho\int_{nT}^{t}\mathbb{E}\left[\left|\bar{\theta}_s^{\lambda} - \bar{\zeta}_s^{\lambda,n}\right|^2\right]\mathrm{d}s$$
$$+ 4e^{-a\lambda nT/2}\lambda(L_1^2C_\rho\bar{\sigma}_Y + \bar{\sigma}_Z)\mathbb{E}[V_2(\theta_0)] + 4\lambda(L_1^2C_\rho\tilde{\sigma}_Y + \tilde{\sigma}_Z).$$

Using Grönwall's inequality leads

$$\mathbb{E}\left[\left|\bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^{\lambda}\right|^2\right] \le \lambda e^{4L_1^2C_\rho}\left[4e^{-a\lambda nT/2}(L_1^2C_\rho\bar{\sigma}_Y + \bar{\sigma}_Z)\mathbb{E}[V_2(\theta_0)] + 4(L_1^2C_\rho\tilde{\sigma}_Y + \tilde{\sigma}_Z)\right].$$

which implies by $\lambda T \ge 1/2$,

$$\tilde{W}_2^2(\mathcal{L}(\bar{\theta}_t^{\lambda}), \mathcal{L}(\bar{\zeta}_t^{\lambda,n})) \le \mathbb{E}\left|\bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^{\lambda}\right|^2 \le \lambda(e^{-an/4}\bar{C}_{2,1}\mathbb{E}[V_2(\theta_0)] + \bar{C}_{2,2}),$$

where

$$\bar{C}_{2,1} = 4e^{4L_1^2C_\rho}(L_1^2C_\rho\bar{\sigma}_Y + \bar{\sigma}_Z), \quad \bar{C}_{2,2} = 4e^{4L_1^2C_\rho}(L_1^2C_\rho\tilde{\sigma}_Y + \tilde{\sigma}_Z) \tag{26}$$

with $\bar{\sigma}_Y$, $\tilde{\sigma}_Y$ provided in (31) and $\bar{\sigma}_Z$, $\tilde{\sigma}_Z$ given in (30). ∎

Lemma 3.8. *Let Assumption 1, 2 and 3 hold. For any* $0 < \lambda < \lambda_{\max}$ *given in* (8), $t \in [nT, (n+1)T]$,

$$W_1(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}), \mathcal{L}(Z_t^{\lambda})) \le \sqrt{\lambda}(e^{-\min\{\dot{c},a/4\}n/2}\bar{C}_{2,3}\mathbb{E}[V_4(\theta_0)] + \bar{C}_{2,4}),$$

where $\bar{C}_{2,3}$*,* $\bar{C}_{2,4}$ *is given in* (27)*.*

Proof. To upper bound the second term $W_1(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}), \mathcal{L}(Z_t^{\lambda}))$ in (23), we adapt the proof from Lemma 3.28 in [5]. By Proposition 3.6, Corollary 3.3, Lemma 3.5 and 3.7, one obtains

$$\begin{aligned}
&W_1(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}), \mathcal{L}(Z_t^{\lambda}))\\
&\le \sum_{k=1}^{n} W_1(\mathcal{L}(\bar{\zeta}_t^{\lambda,k}), \mathcal{L}(\bar{\zeta}_t^{\lambda,k-1})),\\
&\le \sum_{k=1}^{n} w_{1,2}(\mathcal{L}(\zeta_t^{kT,\bar{\theta}_{kT}^{\lambda},\lambda}), \mathcal{L}(\zeta_t^{kT,\bar{\zeta}_{kT}^{\lambda,k-1},\lambda}))\\
&\le \hat{c}\sum_{k=1}^{n}\exp(-\dot{c}(n-k))w_{1,2}(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}), \mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1}))\\
&\le \hat{c}\sum_{k=1}^{n}\exp(-\dot{c}(n-k))\tilde{W}_2(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}), \mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1}))\left[1 + \left\{\mathbb{E}[V_4(\bar{\theta}_{kT}^{\lambda})]\right\}^{1/2} + \left\{\mathbb{E}[V_4(\bar{\zeta}_{kT}^{\lambda,k-1})]\right\}^{1/2}\right]\\
&\le (\sqrt{\lambda})^{-1}\hat{c}\sum_{k=1}^{n}\exp(-\dot{c}(n-k))\tilde{W}_2^2(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}), \mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1}))\\
&\quad + 3\sqrt{\lambda}\hat{c}\sum_{k=1}^{n}\exp(-\dot{c}(n-k))\left[1 + \mathbb{E}[V_4(\bar{\theta}_{kT}^{\lambda})] + \mathbb{E}[V_4(\bar{\zeta}_{kT}^{\lambda,k-1})]\right]\\
&\le \sqrt{\lambda}e^{-\min\{\dot{c},a/4\}n}n\hat{c}(e^{\min\{\dot{c},a/4\}}\bar{C}_{2,1}\mathbb{E}[V_2(\theta_0)] + 12\mathbb{E}[V_4(\theta_0)])
\end{aligned}$$

$$
\begin{aligned}
&+ \sqrt{\lambda}\frac{\hat{c}}{1-\exp(-\dot{c})}(\bar{C}_{2,2} + 12c_3(\lambda_{\max} + a^{-1}) + 9v_4(\bar{M}_4) + 15) \\
&\leq \sqrt{\lambda}(e^{-\min\{\dot{c},a/4\}n/2}\bar{C}_{2,3}\mathbb{E}[V_4(\theta_0)] + \bar{C}_{2,4})
\end{aligned}
$$

where the last inequality holds due to $e^{-\alpha n}(n+1) \leq 1+\alpha^{-1}$, for $\alpha > 0$, and we take $\alpha = \min\{\dot{c}, a/4\}/2$, moreover,

$$
\begin{aligned}
\bar{C}_{2,3} &= \hat{c}\left(1 + \frac{2}{\min\{\dot{c}, a/4\}}\right)(e^{\min\{\dot{c},a/4\}}\bar{C}_{2,1} + 12) \\
\bar{C}_{2,4} &= \frac{\hat{c}}{1-\exp(-\dot{c})}(\bar{C}_{2,2} + 12c_3(\lambda_{\max} + a^{-1}) + 9v_4(\bar{M}_4) + 15)
\end{aligned}
\tag{27}
$$

with $\bar{C}_{2,1}$, $\bar{C}_{2,2}$ given in 26, $\hat{c}$, $\dot{c}$ given in Lemma 3.10, c_3 is given in (15) and $\bar{M}_4$ given in Lemma 3.4. □

Lemma 3.9. *Let Assumption 1, 2 and 3 hold. For any $0 < \lambda < \lambda_{\max}$ given in (8), $t \in [nT, (n+1)T]$,*

$$
W_1(\mathcal{L}(\bar{\theta}_t^\lambda), \mathcal{L}(Z_t^\lambda)) \leq \bar{C}_2\sqrt{\lambda}(e^{-\min\{\dot{c},a/4\}n/2}\mathbb{E}[V_4(\theta_0)] + 1),
$$

where $\bar{C}_2$ is given in (28).

Proof. By using Lemma 3.7 and 3.8, one obtains

$$
\begin{aligned}
&W_1(\mathcal{L}(\bar{\theta}_t^\lambda), \mathcal{L}(Z_t^\lambda)) \\
&\leq W_1(\mathcal{L}(\bar{\theta}_t^\lambda), \mathcal{L}(\bar{\zeta}_t^{\lambda,n})) + W_1(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}), \mathcal{L}(Z_t^\lambda)) \\
&\leq \sqrt{\lambda}(e^{-an/8}\bar{C}_{2,1}^{1/2}\mathbb{E}^{1/2}[V_2(\theta_0)] + \bar{C}_{2,2}^{1/2}) + \sqrt{\lambda}(e^{-\min\{\dot{c},a/4\}n/2}\bar{C}_{2,3}\mathbb{E}[V_4(\theta_0)] + \bar{C}_{2,4}) \\
&\leq \bar{C}_2\sqrt{\lambda}(e^{-\min\{\dot{c},a/4\}n/2}\mathbb{E}[V_4(\theta_0)] + 1),
\end{aligned}
$$

where

$$
\bar{C}_2 = \bar{C}_{2,1}^{1/2} + \bar{C}_{2,2}^{1/2} + \bar{C}_{2,3} + \bar{C}_{2,4}. \tag{28}
$$

□

Before proceeding to the proofs of the main results, we provide explicitly the constants $\dot{c}$ and $\hat{c}$ in Proposition 3.6.

Lemma 3.10. *The contraction constant in Proposition 3.6 is given by*

$$
\dot{c} = \min\{\bar{\phi}, \bar{c}(p), 4\tilde{c}(p)\epsilon\bar{c}(p)\}/2,
$$

where the explicit expressions for $\bar{c}(p)$ and $\tilde{c}(p)$ can be found in Lemma 3.4 and $\bar{\phi}$ is given by

$$
\bar{\phi} = \left(\sqrt{4\pi/K_1}\bar{b}\exp\left(\left(\bar{b}\sqrt{K_1}/2 + 2/\sqrt{K_1}\right)^2\right)\right)^{-1}.
$$

Furthermore, any ϵ can be chosen which satisfies the following inequality

$$
\epsilon \leq 1 \wedge \left(8\tilde{c}(p)\sqrt{\pi/K_1}\int_0^{\tilde{b}}\exp\left(\left(s\sqrt{K_1}/2 + 2/\sqrt{K_1}\right)^2\right)\,\mathrm{d}s\right)^{-1},
$$

where $K_1 = L_1\mathbb{E}[(1+|X_0|)^\rho]$, $\tilde{b} = \sqrt{2\tilde{c}(p)/\bar{c}(p) - 1}$ and $\bar{b} = \sqrt{4\tilde{c}(p)(1+\bar{c}(p))/\bar{c}(p) - 1}$. The constant $\hat{c}$ is given as the ratio C_{11}/C_{10}, where C_{11}, C_{10} are given explicitly in [5, Lemma 3.26].

Proof. See [5, Lemma 3.26]. □

Proof of Theorem 2.3 One notes that, by Lemma 3.9, for $t \in [nT, (n+1)T]$

$$\begin{aligned} W_1(\mathcal{L}(\theta_t^\lambda), \pi_\beta) &\le W_1(\mathcal{L}(\bar{\theta}_t^\lambda), \mathcal{L}(Z_t^\lambda)) + W_1(\mathcal{L}(Z_t^\lambda), \pi_\beta) \\ &\le \bar{C}_2\sqrt{\lambda}(e^{-\min\{\dot{c}, a/4\}n/2}\mathbb{E}[V_4(\theta_0)] + 1) + \hat{c}e^{-\dot{c}\lambda t} w_{1,2}(\theta_0, \pi_\beta) \\ &\le \bar{C}_2\sqrt{\lambda}(e^{-\min\{\dot{c}, a/4\}n/2}\mathbb{E}[V_4(\theta_0)] + 1) + \hat{c}e^{-\dot{c}\lambda t}\left[1 + \mathbb{E}[V_2(\theta_0)] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right] \\ &\le 2e^{-\min\{\dot{c}, a/4\}n/2}(\lambda_{\max}^{1/2}\bar{C}_2 + \hat{c})(1 + \mathbb{E}[|\theta_0|^4]) \\ &\quad + \hat{c}e^{-\min\{\dot{c}, a/4\}n/2}\left[1 + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right] + \sqrt{\lambda}\bar{C}_2, \end{aligned}$$

which implies, for any $n \in \mathbb{N}$

$$W_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \le C_1 e^{-C_0\lambda n}(1 + \mathbb{E}[|\theta_0|^4]) + C_2\sqrt{\lambda},$$

where

$$C_0 = \min\{\dot{c}, a/4\}/2, \quad C_1 = 2\left[(\lambda_{\max}^{1/2}\bar{C}_2 + \hat{c}) + \hat{c}\left(1 + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right)\right], \quad C_2 = \bar{C}_2, \tag{29}$$

with $\bar{C}_2$ given in 28.

Proof of Theorem 2.5 We prove this result without providing explicit constants. By using [11, Corollary 2], the contraction result in Proposition 3.6 can be established in $\tilde{W}_1$ distance instead of $w_{1,2}$. Then, one obtains by noticing $\tilde{W}_1 \le \tilde{W}_2$,

$$\tilde{W}_1(\mathcal{L}(\bar{\theta}_t^\lambda), \mathcal{L}(Z_t^\lambda)) \le \tilde{C}_2\sqrt{\lambda}.$$

Finally, by using the same arugments as in the Proof of Theorem 2.3, one obtains

$$\tilde{W}_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \le C_4 \mathrm{e}^{-C_3\lambda n}\mathbb{E}[|\theta_0|^4 + 1] + C_5\sqrt{\lambda}$$

with appropriate constants $C_3, C_4, C_5 > 0$.

4 Applications

Bayesian logistic regression We consider a sampling problem where the target distribution is the posterior in a Bayesian inference problem with Gaussian mixture distribution prior. One observes a sequence of i.i.d. sample $\{(x_i, y_i)\}_{i=1,\dots,n}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{0,1\}$ for all i. Denote by $\mathbf{x}_i = (x_i, y_i)$ for all i, the likelihood function is given by $p(\mathbf{x}_i|\theta) = p(y_i|x_i, \theta) = (1/(1 + e^{-x_i^\mathsf{T}\theta}))^{y_i}(1 - 1/(1 + e^{-x_i^\mathsf{T}\theta}))^{1-y_i}$, for $\theta \in \mathbb{R}^d$. The Gaussian mixture prior has the form

$$\pi_0(\theta) \propto \exp(-f(\theta)) = e^{-|\theta - \hat{a}|^2/2} + e^{-|\theta + \hat{a}|^2/2}$$

where $\hat{a} \in \mathbb{R}^d$ and $f(\theta) = |\theta - \hat{a}|^2/2 - \log(1 + \exp(-2\hat{a}^\mathsf{T}\theta))$. We choose $\hat{a} \in \mathbb{R}^d$ such that $|\hat{a}|^2 > 1$ for the function f to be nonconvex, see [7] for more discussions. In this case, the stochastic gradient $H(\theta, \mathbf{x}) = -\nabla \log \pi(\theta, \mathbf{x})$ with $\pi(\theta, \mathbf{x}) = \pi_0(\theta)\prod_{i=1}^n p(\mathbf{x}_i|\theta)$, and it can be expressed as

$$H(\theta, \mathbf{x}) = \sum_{i=1}^n H(\theta, \mathbf{x}_i) = \sum_{i=1}^n \left(\frac{1}{n}\left(\theta - \hat{a} + \frac{2\hat{a}}{1 + e^{2\hat{a}^\mathsf{T}\theta}}\right) + \frac{x_i}{1 + e^{-x_i^\mathsf{T}\theta}} - y_i x_i\right).$$

Then, one notices that, for each $i = 1, \dots, n$

$$\begin{aligned} |H(\tilde{\theta}, \mathbf{x}_i) - H(\hat{\theta}, \mathbf{x}_i)| &= \left|\frac{1}{n}\left(\tilde{\theta} - \hat{\theta} + \frac{2\hat{a}}{1 + e^{2\hat{a}^\mathsf{T}\tilde{\theta}}} - \frac{2\hat{a}}{1 + e^{2\hat{a}^\mathsf{T}\hat{\theta}}}\right) + \left(\frac{x_i}{1 + e^{-x_i^\mathsf{T}\tilde{\theta}}} - \frac{x_i}{1 + e^{-x_i^\mathsf{T}\hat{\theta}}}\right)\right| \\ &\le \left(1 + 4|\hat{a}|^2 + |\mathbf{x}_i|^2\right)|\tilde{\theta} - \hat{\theta}| \\ &\le (1 + 4|\hat{a}|^2)(1 + |\mathbf{x}_i|)^2|\tilde{\theta} - \hat{\theta}|, \end{aligned}$$

and

$$
\begin{aligned}
|H(\theta,\tilde{\mathbf{x}}_i)-H(\theta,\hat{\mathbf{x}}_i)| &= \left|\left(\frac{\tilde{x}_i}{1+e^{-\tilde{x}_i^\mathsf{T}\theta}}-\frac{\hat{x}_i}{1+e^{-\hat{x}_i^\mathsf{T}\theta}}-(\tilde{y}_i\tilde{x}_i-\hat{y}_i\hat{x}_i)\right)\right| \\
&\le |\tilde{x}_i-\hat{x}_i|+|\tilde{x}_i||e^{-\hat{x}_i^\mathsf{T}\theta}-e^{-\tilde{x}_i^\mathsf{T}\theta}|+|\tilde{x}_i-\hat{x}_i|+|\tilde{y}||\tilde{x}_i-\hat{x}_i|+|\hat{x}||\tilde{y}_i-\hat{y}_i| \\
&\le (2+|\theta||\tilde{\mathbf{x}}_i|+|\tilde{\mathbf{x}}_i|+|\hat{\mathbf{x}}_i|)|\tilde{\mathbf{x}}_i-\hat{\mathbf{x}}_i| \\
&\le 2(1+|\tilde{\mathbf{x}}_i|+|\hat{\mathbf{x}}_i|)^2(1+|\theta|)|\tilde{\mathbf{x}}_i-\hat{\mathbf{x}}_i|
\end{aligned}
$$

which suggests Assumption 2 holds with $\rho=2$, $L_1=1+4|\hat{a}|^2$ and $L_2=2$. To see Assumptions 3 is satisfied, one calculates, for $i=1,\dots,n$,

$$
\begin{aligned}
\langle\theta,H(\theta,\mathbf{x}_i)\rangle &= \frac{1}{n}\left(|\theta|^2+\frac{1-e^{2\hat{a}^\mathsf{T}\theta}}{1+e^{2\hat{a}^\mathsf{T}\theta}}\langle\theta,\hat{a}\rangle\right)+\left(\frac{\langle x_i,\theta\rangle}{1+e^{-x_i^\mathsf{T}\theta}}-y_i\langle x_i,\theta\rangle\right) \\
&\ge \frac{1}{2n}|\theta|^2-\frac{1}{2n}|\hat{a}|^2+\left(-2n|x_i|^2-2n|y_ix_i|^2-\frac{1}{4n}|\theta|^2\right) \\
&\ge \frac{1}{4n}|\theta|^2-\frac{1}{2n}|\hat{a}|^2-2n\left(|\mathbf{x}_i|^2+|\mathbf{x}_i|^4\right),
\end{aligned}
$$

which implies that $A(\mathbf{x}_i)=\mathbf{I}_d/(4n)$ and $b(\mathbf{x}_i)=|\hat{a}|^2/(2n)+2n(|\mathbf{x}_i|^2+|\mathbf{x}_i|^4)$. The smallest eigenvalue of $\mathbb{E}(A(\mathbf{x}_i))$ and $b=\mathbb{E}(b(\mathbf{x}_i))$ can be obtained accordingly.

5 Appendix

Lemma 5.1. *Let Assumption 1, 2 and 3 hold. For any $t\in[nT,(n+1)T]$, $n\in\mathbb{N}$ and $k=1,\dots,K+1$, $K+1\le T$, one obtains*

$$
\mathbb{E}\left[\left|h(\bar{\zeta}_t^{\lambda,n})-H(\bar{\zeta}_t^{\lambda,n},X_{nT+k})\right|^2\right]\le e^{-a\lambda t/2}\bar{\sigma}_Z\mathbb{E}[V_2(\theta_0)]+\tilde{\sigma}_Z,
$$

where

$$
\bar{\sigma}_Z=8L_2^2\hat{\sigma},\quad \tilde{\sigma}_Z=8L_2^2\hat{\sigma}(3\mathrm{v}_2(\overline{M}_2)+c_1(\lambda_{\max}+a^{-1})+1), \tag{30}
$$

with $\hat{\sigma}_Z=\mathbb{E}[(1+|X_0|+|\mathbb{E}[X_0]|)^{2\rho}|X_0-\mathbb{E}[X_0]|^2]$.

Proof. Recall $\mathcal{H}_t=\mathcal{F}_\infty^\lambda\vee\mathcal{G}_{\lfloor t\rfloor}$. One notices that

$$
\begin{aligned}
&\mathbb{E}\left[\left|h(\bar{\zeta}_t^{\lambda,n})-H(\bar{\zeta}_t^{\lambda,n},X_{nT+k})\right|^2\right] \\
&=\mathbb{E}\left[\mathbb{E}\left[\left|h(\bar{\zeta}_t^{\lambda,n})-H(\bar{\zeta}_t^{\lambda,n},X_{nT+k})\right|^2\middle|\mathcal{H}_{nT}\right]\right] \\
&=\mathbb{E}\left[\mathbb{E}\left[\left|\mathbb{E}\left[H(\bar{\zeta}_t^{\lambda,n},X_{nT+k})\middle|\mathcal{H}_{nT}\right]-H(\bar{\zeta}_t^{\lambda,n},X_{nT+k})\right|^2\middle|\mathcal{H}_{nT}\right]\right] \\
&\le 4\mathbb{E}\left[\mathbb{E}\left[\left|H(\bar{\zeta}_t^{\lambda,n},X_{nT+k})-H(\bar{\zeta}_t^{\lambda,n},\mathbb{E}\left[X_{nT+k}\middle|\mathcal{H}_{nT}\right])\right|^2\middle|\mathcal{H}_{nT}\right]\right] \\
&\le 4L_2^2\hat{\sigma}_Z\mathbb{E}\left[\left(1+\left|\bar{\zeta}_t^{\lambda,n}\right|\right)^2\right],
\end{aligned}
$$

where the first inequality holds due to Lemma 5.3 and $\hat{\sigma}_Z=\mathbb{E}[(1+|X_0|+|\mathbb{E}[X_0]|)^{2\rho}|X_0-\mathbb{E}[X_0]|^2]$. Then, by using Lemma 3.5, one obtains

$$
\mathbb{E}\left[\left|h(\bar{\zeta}_t^{\lambda,n})-H(\bar{\zeta}_t^{\lambda,n},X_{nT+k})\right|^2\right]\le 8L_2^2\hat{\sigma}_Z\mathbb{E}\left[V_2(\bar{\zeta}_t^{\lambda,n})\right]\le e^{-a\lambda t/2}\bar{\sigma}_Z\mathbb{E}[V_2(\theta_0)]+\tilde{\sigma}_Z,
$$

where $\bar{\sigma}_Z=8L_2^2\hat{\sigma}$ and $\tilde{\sigma}_Z=8L_2^2\hat{\sigma}(3\mathrm{v}_2(\overline{M}_2)+c_1(\lambda_{\max}+a^{-1})+1)$. ◻

Lemma 5.2. *Let Assumption 1, 2 and 3 hold. For any* $t > 0$, *one obtains*

$$\mathbb{E}\left[\left|\bar{\theta}^{\lambda}_t - \bar{\theta}^{\lambda}_{\lfloor t \rfloor}\right|^2\right] \leq \lambda(e^{-a\lambda\lfloor t \rfloor}\bar{\sigma}_Y \mathbb{E}[V_2(\theta_0)] + \tilde{\sigma}_Y),$$

where

$$\bar{\sigma}_Y = 2\lambda_{\max}L_1^2 C_\rho, \quad \tilde{\sigma}_Y = 2\lambda_{\max}L_1^2 C_\rho c_1(\lambda_{\max} + a^{-1}) + 4\lambda_{\max}L_2^2 C_\rho + 4\lambda_{\max}H_*^2 + 2d\beta^{-1}. \tag{31}$$

Proof. For any $t > 0$, we write the difference $\left|\bar{\theta}^{\lambda}_t - \bar{\theta}^{\lambda}_{\lfloor t \rfloor}\right|$ and use $(a+b)^2 \leq 2a^2 + 2b^2$ which yields

$$\begin{aligned}
\mathbb{E}\left[\left|\bar{\theta}^{\lambda}_t - \bar{\theta}^{\lambda}_{\lfloor t \rfloor}\right|^2\right] &= \mathbb{E}\left[\left|-\lambda\int_{\lfloor t \rfloor}^{t} H(\bar{\theta}^{\lambda}_{\lfloor t \rfloor}, X_{\lceil t \rceil})\mathrm{d}s + \sqrt{\frac{2\lambda}{\beta}}(\tilde{B}^{\lambda}_t - \tilde{B}^{\lambda}_{\lfloor t \rfloor})\right|^2\right] \\
&\leq \lambda^2 \mathbb{E}\left[\left(L_1(1+|X_{\lceil t \rceil}|)^\rho |\bar{\theta}^{\lambda}_{\lfloor t \rfloor}| + L_2(1+|X_{\lceil t \rceil}|)^{\rho+1} + H_*\right)^2\right] + 2d\lambda\beta^{-1},
\end{aligned}$$

where the inequality holds due to Remark 2.1 and by applying Lemma 3.2, one obtains

$$\begin{aligned}
\mathbb{E}\left[\left|\bar{\theta}^{\lambda}_t - \bar{\theta}^{\lambda}_{\lfloor t \rfloor}\right|^2\right] &\leq 2\lambda^2 L_1^2 \mathbb{E}[(1+|X_0|)^{2\rho}]\mathbb{E}[|\bar{\theta}^{\lambda}_{\lfloor t \rfloor}|^2] + 4\lambda^2 L_2^2 \mathbb{E}[(1+|X_0|)^{2\rho+2}] + 4\lambda^2 H_*^2 + 2d\lambda\beta^{-1} \\
&\leq \lambda((1-a\lambda)^{\lfloor t \rfloor}\bar{\sigma}_Y \mathbb{E}[V_2(\theta_0)] + \tilde{\sigma}_Y),
\end{aligned}$$

where $\bar{\sigma}_Y = 2\lambda_{\max}L_1^2 C_\rho$ and $\tilde{\sigma}_Y = 2\lambda_{\max}L_1^2 C_\rho c_1(\lambda_{\max} + a^{-1}) + 4\lambda_{\max}L_2^2 C_\rho + 4\lambda_{\max}H_*^2 + 2d\beta^{-1}$. □

Lemma 5.3. *Let* $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ *be sigma-algebras. Let* $p \geq 1$. *Let* X, Y *be* $\mathbb{R}^d$*-valued random vectors in* L^p *such that* Y *is measurable with respect to* $\mathcal{H} \vee \mathcal{G}$. *Then*

$$\mathbb{E}^{1/p}\left[\|X - \mathbb{E}\left[X \mid \mathcal{H} \vee \mathcal{G}\right]\|^p \mid \mathcal{G}\right] \leq 2\mathbb{E}^{1/p}\left[\|X - Y\|^p \mid \mathcal{G}\right].$$

Proof. See [4, Lemma 6.1]. □

References

[1] M. Barkhagen, N. H. Chau, É. Moulines, M. Rásonyi, S. Sabanis and Y. Zhang. On stochastic gradient Langevin dynamics with stationary data streams in the logconcave case. *Preprint*, 2018. arXiv:1812.02709

[2] N. Brosse, A. Durmus and E. Moulines. The promises and pitfalls of stochastic gradient Langevin dynamics. *Advances in Neural Information Processing Systems*, 8268-8278, 2018.

[3] N. Brosse, A. Durmus, E. Moulines and S. Sabanis. The Tamed Unadjusted Langevin Algorithm. *Stochastic Processes and their Applications*, 129(10): 3638–3663, 2019

[4] N. H. Chau, Ch. Kumar, M. Rásonyi and S. Sabanis. On fixed gain recursive estimators with discontinuity in the parameters. *ESAIM Probability and Statistics*, 23:217–244, 2019.

[5] N. H. Chau, É. Moulines, M. Rásonyi, S. Sabanis and Y. Zhang. On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *Preprint*, 2019. arXiv:1905.13142

[6] X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett and M. I. Jordan. Sharp convergence rated for Langevin dynamics in the nonconvex setting. *Preprint*, 2018. arXiv:1805.01648v1

[7] A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79:651–676, 2017.

[8] A. S. Dalalyan and A. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *To appear in Stochastic Processes and their Applications*, 2019.

[9] A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27:1551–1587, 2017.

[10] A. Durmus and É. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Preprint*, 2018. arXiv:1605.01559v3

[11] A. Eberle. Reflection couplings and contraction rates for diffusions. *Probab. Theory Related Fields*, 166:851–886, 2016.

[12] A. Eberle, A. Guillin and R. Zimmer. Quantitative Harris-type theorems for diffusions and McKean-Vlasov processes. *In Press, Transactions of the American Mathematical Society*, 2018. https://doi.org/10.1090/tran/7576

[13] C.R. Hwang Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, 8(6): 1177-1182, 1980.

[14] M. B. Majka, A. Mijatović and Ł. Szpruch. Non-asymptotic bounds for sampling algorithms without log-concavity. *Preprint*, 2018. arXiv:1808.07105v1

[15] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. *Proceedings of Machine Learning Research*, (65)1674–1703, 2017.

[16] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. *Proceedings of the 28th international conference on machine learning* (ICML-11). 2011.

[17] S. Sotirios and Y. Zhang. Higher Order Langevin Monte Carlo Algorithm. *Electronic Journal of Statistics*, 13(2): 3805-3850, 2019.